Group 3: Jeannie He, Li Cheng, Yifei Dong, Yusen Wang

HW3.1 For $f^*(y) = \max_x y^T x - f(x)$

There is: $\partial f^*(y) = x^* = \arg\max_x y^T x - f(x)$

$= \arg\min_x f(x) - y^T x$. ①

For a closed and convex f, it can be proved that:

$\partial g(\lambda) = A^T \partial f^*(-A^T\lambda) - b$ (affine transformations of domain)

According to ① with $y = -A^T\lambda$, there is:

$\partial f^*(-A^T\lambda) = \arg\min_w [f(w) + \lambda^T A w]$

$= \arg\min_w [f(w) + \lambda^T A w - \lambda^T b]$

$= w^*$

$\therefore Aw - b \in \partial g(\lambda)$. Q.E.D.

H3.2. Here we first prove that, f is μ-strongly convex $\Leftrightarrow$ f^* is $\frac{1}{\mu}$-smooth.
Then according to $f^{**} = f$, f is L-smooth $\Leftrightarrow$ f^* is $\frac{1}{L}$-strongly convex.
Finally for the $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth f^*, prove the convergence of dual ascent.

① Prove f is μ-strongly convex $\Leftrightarrow$ f^* is $\frac{1}{\mu}$-smooth.

Proof: Let $Z = f - y^T x$, so Z is also μ-strongly convex.

of "$\Rightarrow$" Defining $x_u = \nabla f^*(u)$, $x_v = \nabla f^*(v)$, For a minimizer x^*, there are:

$$Z(x) \geq Z(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2 \quad \text{since } \nabla Z(x^*) = 0.$$

So:
$$\begin{cases} f(x_v) - u^T x_v \geq f(x_u) - u^T x_u + \frac{\mu}{2}\|x_u - x_v\|_2^2 \\ f(x_u) - v^T x_u \geq f(x_v) - v^T x_v + \frac{\mu}{2}\|x_u - x_v\|_2^2 \end{cases}$$

Adding them together: $(u^T - v^T)(x_u - x_v) \geq \mu\|x_u - x_v\|_2^2$

With Cauchy-Schwartz: $(u^T - v^T)(x_u - x_v) \leq \|u^T - v^T\|\,\|x_u - x_v\|$

So $\|x_u - x_v\| \leq \|u - v\|/\mu$

$\Rightarrow$ ∇f^* is Lipschitz with $\frac{1}{\mu}$ $\Rightarrow$ f^* is $\frac{1}{\mu}$-smooth.

Proof of "$\Leftarrow$": For $\frac{1}{\mu}$-smooth f^*, ~~there is: let $g_x(y) = f(y) - \nabla g(x)^T y$.~~

let $g_x(y) = f^*(y) - \nabla f^*(x)^T y$.

$$\text{so} \quad g_x(y) \leq g_x(y) + \nabla g_x(y)^T(y - y) + \frac{1}{2\mu}\|y - y\|_2^2.$$

Minimizing each side over y, and rearranging, there is:

$$\frac{\mu}{2}\|\nabla f^*(x) - \nabla f^*(y)\|_2^2 \leq f^*(y) - f^*(x) + \nabla f^*(x)^T(x - y).$$

Exchange x and y: $\frac{\mu}{2}\|\nabla f^*(x) - \nabla f^*(y)\|_2^2 \leq f^*(x) - f^*(y) + \nabla f^*(y)^T(y - x)$

Adding together: $\mu\|\nabla f^*(x) - \nabla f^*(y)\|_2^2 \leq [\nabla f^*(x) - \nabla f^*(y)]^T(x - y)$

Let $u = \nabla f^*(x)$, $v = \nabla f^*(y)$, so $(x - y)^T(u - v) \geq \mu\|u - v\|_2^2$

With the proof of "$\Leftarrow$" and "$\Rightarrow$", Q.E.D.

② Study the convergence of dual ascent.

For $\frac{1}{\mu}$-smooth f^*, a fixed step $\alpha = 1/\frac{1}{\mu} = \mu$ is often selected.

$$x_{k+1} = x_k + \mu \nabla f^*(x_k)$$

$$\because f^*(y) \le f^*(x) + \nabla f^*(x)(y-x) + \frac{1}{2\mu}\|y-x\|_2^2$$

$$f^*(x_{k+1}) \le f^*(x_k) + \frac{3\mu}{2}\|\nabla f^*(x_k)\|_2^2 \quad ①$$

$$\Rightarrow f^*(x_k) - f^*(x^*) \le \frac{3\mu}{2k}\|x_0 - x^*\|_2^2$$

to guarantee $f^*(x_k) - f^*(x^*) \le \epsilon$, there should be $k \ge \frac{3\mu}{2\epsilon}\|x_0 - x^*\|_2^2$

$\therefore k \sim O(\frac{1}{\epsilon})$ sublinear.

For $\frac{1}{L}$-strongly convex f^*, there are:

$$f^*(y) \ge f^*(x) + \nabla f^*(x)(y-x) + \frac{1}{2L}\|y-x\|_2^2.$$

Minimize this lower bound by taking the gradient w.r.t y and setting it to 0:

$$\nabla f^*(x) + \frac{1}{L}(y-x) = 0.$$

So the lower bound is: $f^*(y) \ge f^*(x) - L\|\nabla f^*(x)\|_2^2 + \frac{L}{2}\|\nabla f^*(x)\|_2^2$

$$= f^*(x) - \frac{L}{2}\|\nabla f^*(x)\|_2^2$$

$$\therefore \|\nabla f^*(x)\|_2^2 \ge \frac{2}{L}[f^*(x) - f^*(y)]$$

let $y = x^*$ and combined with ①:

$$f^*(x_{k+1}) - f^*(x^*) \le f^*(x_k) - f^*(x^*) + \frac{3\mu}{L}[f^*(x_k) - f^*(x^*)]$$

$$= (1 + \frac{3\mu}{L})[f^*(x_k) - f^*(x^*)].$$

$$\therefore f^*(x_k) - f^*(x^*) \le (1 + \frac{3\mu}{L})^k \cdot [f^*(x_0) - f^*(x^*)] \le \epsilon$$

$$\Rightarrow k \ge \frac{\log[(f^*(x_0) - f^*(x^*))/\epsilon]}{\log[(L+3\mu)/L]} \sim O(\log\frac{1}{\epsilon})$$

linear.

However, the solution is not guaranteed to be primal feasible, since while the dual variables converge to a solution, the primal variables may not satisfy the constraints of the original problem.

HW 3.3

We have the following problem:

$$(P2): \text{minimize } \frac{1}{N}\sum_{i\in[N]} f_i(w_i)$$

$$\text{s.t. } w_i = w_j \text{ for all } j \in \mathcal{N}_i$$

Primal method:

$$\bar{w}_i^k = \sum_{j\in\mathcal{N}_i} a_{ij} w_j^k \quad \text{(consensus)}$$

$$w_i^{k+1} = \bar{w}_i^k + \alpha_k g_i(\bar{w}_i^k)$$

Dual method:

$$w_i^{k+1} = \underset{w_i}{\arg\min}\, \mathcal{L}_i(w_i, \lambda_i^k) \quad \text{where } \mathcal{L}_i(w_i, \lambda_i^k) = f_i(w_i) + \sum_{i=1}^{N} a_{ij} \lambda_i^{k T}(w_i - w_j)$$

$$\lambda_i^{k+1} = \lambda_i^k + \alpha_k\Big(\sum_{j=1}^{N} a_{ij}(w_j^{k+1} - w_i^{k+1})\Big) \quad \text{(consensus)}$$

Communication cost: Since both the primal and dual method are decentralized with w_i^k being the only variable to be shared, their communication cost per iteration is the same. More specifically, if $w_i^k \in \mathbb{R}^n$ and there is N nodes in $\mathcal{N}_i$ for $i=1,\dots,N$, then the communication cost per iteration is $O(N^2 n)$ in both methods.

Convergence rate: Assuming f is strongly convex with parameter m & L-lipschitz continuous (A1).
Let $f(w_{best}^k) = \min_{i=1,\dots,k} f(w^i)$, $w^* = \lim_{k\to\infty} w^k$, $R = \|w^0 - w^*\|_2$ and $f^* = f(w^*)$. We have

For Primal method: $\|w^k - w^*\|_2^2 \le \|w^{k-1} - w^*\|_2^2 - 2\alpha_k(f(w^{k-1}) - f(w^*)) + \alpha_k^2 \|g(w^{k-1})\|_2^2$

$$\le \|w^0 - w^*\|_2^2 - 2\sum_{i=1}^{k}\alpha_i(f(w^{i-1}) - f(w^*)) + \sum_{i=1}^{k}\alpha_k^2\|g(w^{i-1})\|_2^2$$

$$\Rightarrow 0 \le \|w^k - w^*\|_2^2 \le R^2 - 2\sum_{i=1}^{k}\alpha_i(f(w^{i-1}) - f(w^*)) + \sum_{i=1}^{k}\alpha_k^2\|g(w^{i-1})\|_2^2$$

Under assumption A1: $\lim_{k\to\infty} f(w_{best}^k) \le f(w^*) + L^2\alpha/2$

$$\Rightarrow f(w_{best}^k) - f(w^*) \le \frac{R^2 + L^2\sum_{i=1}^{k}\alpha_k^2}{2\sum_{i=1}^{k}\alpha_k}$$

Next page

HW 3.3 continued:

For simplicity, let $\alpha_k = \alpha$ for $k = 1, 2, \ldots$. Then:

$$f(w^*_{best}) - f(w^*) \leq \frac{R^2 + L^2 \sum_{\tau=1}^{k} \alpha_k^2}{2 \sum_{\tau=1}^{k} \alpha_k} = \frac{R^2 + L^2 k \alpha^2}{2k\alpha} \leq \varepsilon \text{ when } \frac{R^2 + L^2 k \alpha^2}{2k\alpha} \leq \varepsilon$$

If we choose α so that $R^2 = L^2 k \alpha^2$, then the above holds when $\frac{R^2}{2k\alpha} = \frac{L^2\alpha}{2} \leq \frac{\varepsilon}{2}$

$\Rightarrow \alpha \leq \frac{\varepsilon}{L^2}$ and $k \geq \frac{R^2}{\alpha\varepsilon} = \frac{R^2 L^2}{\varepsilon^2}$

Hence: The primal method has convergence rate $O\left(\frac{1}{\varepsilon^2}\right)$

Let now f^* be the conjugate function of f and $w_x = \nabla f^*(x)$

Under assumption A1, we have: $f(w_x) \geq f(w_y) + \frac{m}{2} \| w_x - w_y \|_2^2$

$$\Rightarrow \begin{cases} f(w_x) - y^T w_x \geq f(w_y) - y^T w_y + \frac{m}{2} \| w_x - w_y \|_2^2 \\ f(w_y) - x^T w_y \geq f(w_x) - x^T w_x + \frac{m}{2} \| w_y - w_x \|_2^2 \end{cases}$$

Adding these gives:

$\Rightarrow \| \nabla f^*(x) - \nabla f^*(y) \|_2 \leq \frac{1}{m} \| x - y \|_2 = L \| x - y \|_2$ if $L = \frac{1}{m}$

By applying the properties of gradient descent and the fact that the dual method is about solving the dual problem by minimizing the Lagrange function corresponding to maximizing $-f^*(\cdot)$, we see thus that the convergence rate of the dual method is $O(\log(\frac{1}{\varepsilon}))$ if we choose $\alpha = \frac{2}{\left(\frac{1}{m} + \frac{1}{L}\right)}$.

This shows that the dual method has a faster convergence rate than the primal method.